1985

# USDA/SRS SOFTWARE FOR LANDSAT MSS-BASED CROP-ACREAGE ESTIMATION

by Martin Ozga
Statistical Reporting Service
United States Department of Agriculture

## 1. INTRODUCTION

In its continuing development of crop acreage estimation using Landsat MSS data, the Statistical Reporting Service (SRS) of the United States Department of Agriculture (USDA) has developed a large software system known as EDITOR. EDITOR has been developed both internally and with the aid of other agencies and also universities. EDITOR is a large collection of independent programs operating on data in many different files. Most EDITOR programs are called using relatively simple commands from a main program.

An important feature of EDITOR has always been that some programs may run on machines other than that used for most of the processing. For example, supercomputers have long been used by SRS for full-scene processing and even some smaller-scale processing. More recently, microcomputers are being used to do segment boundary digitization. With the constantly growing assortment of machines available, this distribution of functions can be expected to increase.

## 2. HISTORY AND DEVELOPMENT OF EDITOR

EDITOR was initially intended to apply certain functions of LARSYS [1] to the ILLIAC-IV [2], an early supercomputer developed at the University of Illinois. LARSYS is a Landsat MSS processing system developed at the Laboratory for Applications of Remote Sensing (LARS) at Purdue University. Development of EDITOR began at the Center for Advanced Computation (CAC) of the University of Illinois, which at the time (early 1970's) still had some connection to the ILLIAC-IV, even though the ILLIAC-IV had been moved from the University of Illinois to NASA-Ames near Sunnyvale, California. Portions of EDITOR not suitable for the ILLIAC-IV were written for a DECSYSTEM-10 with the TENEX operating system. This system was chosen since it was the front end processor for the ILLIAC-IV. Early funding was from NASA-Ames and the Department of Interior. Later, SRS became interested in EDITOR and changed the course of development to crop acreage estimation--its central focus today.

As development progressed, the EDITOR system was finally moved to the TENEX systems at Bolt Beranek and Newman (BBN) in Cambridge, Massachusetts. BBN provides a reliable time sharing service and also a connection to ARPANET, a Department of Defense research computer network. ARPANET allowed access to both BBN and the ILLIAC-IV from SRS in Washington, D.C., and from CAC in Urbana, Illinois. Due to conditions at the University of Illinois, CAC was discontinued and development work on EDITOR continued at SRS in Washington and also at NASA-Ames. Later, the ILLIAC-IV, a one of a kind machine, was discontinued because it was too expensive to maintain and replaced by a

CRAY 1-S computer [3] later upgraded to a CRAY X-MP. The supercomputer portions of EDITOR were rewritten for the CRAY. Also, BBN discontinued its TENEX systems since they had become obsolete and the bulk of EDITOR was moved to the similar DEC-20 with the TOPS-20 operating system. ARPANET is still used to access the CRAY via a front-end VAX.

Finally, there has recently been an interest in using some EDITOR functions on other machines, including other mainframes, super-minicomputers, and even 32-bit super-microcomputers. During its long development, EDITOR programs were written in several languages and changed often. Therefore, a rewrite of the system was seen as appropriate. PASCAL has been selected as the language for the new Portable EDITOR system, dubbed PEDITOR, with modules being rewritten by both SRS and NASA-Ames. The new PEDITOR is meant to be transportable to many systems. The supercomputer portion of EDITOR, however, must currently remain machine-specific to take advantage of the architecture of such machines.

## 3. ESTIMATION METHOD

Before describing the structure of EDITOR, it is necessary to briefly describe the SRS estimation procedure. The estimates are produced by strata, that is areas of land with similar land-use characteristics. Within each stratum, areas of land known as segments are randomly selected for data collection. The total population of possible segments within a stratum is the number of frame units in the stratum. The process of delineating strata and segments within the strata is known as the construction of an area-sampling frame [4]. The information about land use in the selected segments is collected each year in late May and early June during the June Enumerative Survey (JES) by SRS enumerators who visit the segments and interview the farmers.

Without Landsat MSS, the estimates produced are based on expansions to the entire population of the data collected during the JES. With Landsat MSS, a regression estimator is used with categorized Landsat MSS data as an additional variable to provide current information over the entire area and thus improve the quality of the estimate [5,6]. As EDITOR is currently implemented, estimates using Landsat MSS are done by analysis district, where an analysis district is one Landsat MSS scene or several scenes on the same path taken on the same date. Generation of estimates by county is currently being studied and likely will be made part of the standard EDITOR estimation procedure.

## 4. STRUCTURE AND MAJOR FUNCTIONS OF EDITOR

### 4.1. Overview

EDITOR consists of a large collection of individual programs. Each program is self-contained and may be run separately from any other program. There is no passing of parameters between programs. Typically, an EDITOR program will read in one or more files, process the information, and create one or more output files which may be new files or updated versions of certain of the input files. A few EDITOR programs only display the contents of files.

EDITOR provides a main program allowing the user to call various programs using simple commands. Each program must specify the names of all input and output files. In many cases, particular types of files have standard names so that the user need not enter file names explicitly. This is particularly useful for programs which must process a large number of files of the same type. Jobs requiring use of the CRAY are run in batch mode. An EDITOR program is available, however, to create the input job files required for batch submission to the CRAY. It requires the user to input only a few items, mostly file names.

EDITOR remains very much a system used to generate numbers, that is estimates. It does not provide facilities for display of Landsat MSS data, either raw or categorized, on graphics devices. The only display of Landsat MSS data provided is a grey-scale printout used to help determine the exact location of the land area segments used for ground-truth data. Future plans for use of graphics devices are unclear.

The major files and functions of EDITOR will be described here. The files will be given in approximately the order generally created during the estimation procedure although where there is no dependency, as with the frame unit and segment catalog files, the order of creation may be changed. The functions will also be described in approximate order of use. Again this may vary where there is no dependency. Generally, each function will correspond to a single EDITOR program.

## 4.2. Files Used in EDITOR

The following files are required when using EDITOR for crop-acreage estimation:

1. Segment catalog file, containing various attributes for each segment such as county and stratum.
2. Frame unit file, containing the number of frame units by county and strata.
3. Frame calibration file, containing polynomial coefficients for transformation between map and Landsat MSS coordinates.
4. Segment network or segment video mask, depending on the method of digitization used, containing the digitized field boundaries. For strata in counties, the strata network file contains digitized strata boundaries in counties.
5. Mask file containing an overlay of the segment field or county strata boundaries onto the Landsat MSS coordinate system.
6. Ground data file, one for each segment, containing for each field, the field name, size, cover, etc.
7. Totals file containing a single ground data value for each segment and option selection such as cover type.
8. Landsat MSS data window file containing all the Landsat MSS data for a segment.
9. Packed Landsat MSS file containing only those pixels corresponding to selected ground covers for a group of segments.
10. Statistics file, containing the means and variance-covariance matrix for various categories of Landsat MSS data, usually with one or more categories representing a particular cover.
11. Categorized Landsat MSS data files, both window and packed,

containing a category for each Landsat MSS pixel.

12. Table file, containing a tabulation by segment of pixels by category and reported cover.
13. Estimator parameter file, containing parameters from sample estimation to be used in large-scale estimation.
14. Aggregation file, containing a tabulation, generally for a county, of pixels in each strata and category.
15. Estimator results file, containing the estimates and variances computed by large-scale estimation.

As will be seen, these files are created and read by EDITOR programs. The actual segment catalog, frame unit, and ground data information files are generated outside the scope of EDITOR, but entered into EDITOR files using EDITOR programs.

## 4.3. Segment Catalog File Creation

The segment catalog file is created interactively using information obtained externally to EDITOR. Note that if the same state is processed again in another year, it is not necessary to re-create the segment catalog file, but rather only to update portions of it.

## 4.4. Frame Unit File Creation

The frame unit file is created interactively using information obtained externally to EDITOR. Note that if the same state is processed again in another year, it is not necessary to re-create the frame unit file, but rather only to update portions of it. Automatic procedures also exist to compute the number of frame units in portions of a split county, a county in more than one scene, and to accordingly update the frame unit file.

## 4.5. Landsat MSS Scene Registration

A Landsat MSS scene is registered by finding corresponding points on a picture product of the scene and on topographic maps. For greater accuracy, this is done on a digitizing tablet. The corresponding points are used to generate a least-squares fit to a polynomial. The polynomial is in turn used to check the quality of the points for a smooth fit. The program allows deletion of old points and addition of new ones, until a satisfactory registration is obtained.

Although not yet so implemented, scene registration would seem to be an excellent application for a small microcomputer, since it is highly interactive, does not require too much computation or storage, and generates a small output file which may easily be transferred to a larger machine.

## 4.6. Digitization and Registration of Segments and Counties

Registration of segments and counties, also called calibration, is done on a digitizing tablet using corresponding points to generate a least-squares polynomial mapping between the map and segment or county coordinate systems. This registration may be done on small microcomputers or larger machines and is currently done on both.

Segment digitization is done using one of two methods--manual digitization or video digitization.

For manual digitization, a digitizing tablet is used. Each field is assumed to be a polygon and the vertices are marked on the image. This causes a small, negligible distortion of field boundaries. The fields are digitized by placing the cursor at each vertex and recording the vertex location. The resultant segment network file is a collection of edges and vertices representing the polygon digitized. Manual digitization of segments is done both on microcomputers and at BBN. The microcomputers are favored due to lower cost and generally superior response for this highly interactive task. County digitization is done only at BBN due to memory requirements.

For video digitization [7], the registration of the segment is done on a digitizing tablet as before. However, the field boundaries are traced onto clear acetate. The image is then captured into a raster frame buffer using a video camera. The image is always kept in raster format with no attempt ever made to generate lines or polygons. A connectivity analysis is performed to distinguish the individual fields by connecting all interior raster elements not separated by a boundary element. The boundaries are thinned in the sense of removing excess boundary pixels but never letting previously separate fields be joined. Using the graphics system, the image is displayed and fields are labelled interactively by the user. The result is a video mask, a raster image in the coordinates of the graphics system. Currently, video digitization is performed on a PDP-11 and is available only for segments, due to video resolution restrictions.

### 4.7. Registration of Segments and Counties to the Landsat MSS Coordinate System

To be useful, the segments and counties must be registered to the Landsat MSS coordinate system. This is simply done since polynomials have already been created to transform between Landsat MSS and map coordinates and between segment or county and map coordinates. The manually digitized segment and county network files are converted to a raster representation in the Landsat MSS coordinate system and stored as mask files. The video digitized segments are already in raster format, but in the video coordinate system. They are transformed into the generally coarser Landsat MSS coordinate system to also create mask files. The transformation of video masks is done on the PDP-11.

For counties, the resultant registration is sufficiently precise. However, since segments are much smaller and are used to provide training data, the registration is often not sufficiently accurate due to possible inaccuracies in scene or segment registration. Therefore, segment shifting is required. The shifting is a vertical and/or horizontal shift of the segment against a grey-scale printout to make the segment boundaries more clearly correspond to the Landsat MSS image. An attempt to automate segment shifting using a correlation algorithm has so far been only partially successful. Perhaps a different, more powerful correlation algorithm will be tried in the future on the CRAY.

## 4.8. Editing and Checking of Ground Data

The ground data is prepared outside the scope of the EDITOR system. Typically, a tape containing the ground data information collected by the USDA enumerators is created and shipped to BBN. This tape may have been only partially edited due to timing of follow-up ground data collection and thus may contain errors. The ground truth editor is used to read this tape and create ground data files. The data on the tape are assumed to be in characters or "card-image" form. The user describes the tape with a FORTRAN-like format and a corresponding list of elements. The elements are land cover, field size, etc. Since all data on the tape is numeric, for elements such as land cover, the user must specify a correspondence between the name used in EDITOR and the number on the tape. EDITOR maintains a set of names for each element in a particular file. This file may be updated using a special program not directly callable by EDITOR. The names for land covers, for example, are corn, wheat, etc. The position of the name within the file defines its index for EDITOR processing.

The ground truth editor provides many checks on the ground data files. These checks are both for internal consistency and also checks against the digitized file to make sure both have the same fields of approximately the same size. Also, the ground truth editor provides extensive update capabilities on ground data files.

## 4.9. Creating the Totals File

The totals file represents a summary of certain information from the ground data files and is an input to sample estimation. The SELECT REGION statement is used to specify a list of segments. This may be a list of segments or a Boolean expression of segment attributes used to find a list of segments from the segment catalog file. The value to be entered for each segment is determined by a SELECT OPTIONS statement, a Boolean expression of field attributes used to select fields from the ground data files. The value entered is the sum of the selected size for all fields selected. Once created, the totals file may be updated by changing values, adding new options, or adding new segments.

## 4.10. Packed File Creation

A packed file is created by using a SELECT REGION statement to generate a list of segments and a SELECT OPTIONS statement to select the fields to use in the segments. For each segment, the ground data file is used to determine which fields will be selected based on field attributes. The mask file is used to determine which pixels are copied from the input window file to the output packed file. Typically, a packed file will contain all pixels for a single cover. A packed file containing pixels from all fields with known covers is useful to test classification accuracy. An automatic packing option is available for creating packed files for all covers of interest.

## 4.11. Clustering the Packed Files

Once a packed file is created, it is clustered to obtain a statistics file representative of the cover. The statistics file may contain one or more categories depending on variation in the spectral characteristics of the

cover. Two clustering algorithms are used, the ISODATA algorithm [8] and the more recent CLASSY [9] algorithm. CLASSY is more complex and is generally found to produce better results but has a longer execution time. Therefore, even though it has not been vectorized, CLASSY has been implemented on the CRAY as well as at BBN. Most CLASSY jobs are sent to the CRAY since even its scalar execution times are very fast.

### 4.12. Classifying a Packed File

The packed file for all covers is classified. This is done both to test accuracy of classification and to provide an input to tabulation. The classification may be done either on BBN or the CRAY. The maximum likelihood classification algorithm used is a near-ideal application for the vector architecture of the CRAY so packed file classification is very fast.

### 4.13. Tabulation

Tabulation uses the packed classified file, the ground data file, and the segment mask file to create a table file. The list of segments as well as an encoded representation of the SELECT OPTIONS statement used in creating the packed file are saved in the file header during the packing process. As with packing, the ground data files are used to select the fields to be processed. Then, the mask file is used to get groups of pixels to associate with a field. The group of pixels is assigned the cover of the field, thus allowing a tabulation by cover and category. Raw data tabulation is also of some use. In that case, all pixels are assumed to be in a single category.

### 4.14. Sample Estimation

Sample estimation uses the table file, the totals file, the segment catalog file, and the frame unit file to generate some preliminary estimation results but, more importantly, to create the estimator parameter file as an input to large-scale estimation. The type of estimation to be used is selected during sample estimation. The allowed choices are single-variable regression and multi-variable regression, with single-variable regression the most commonly used. Estimates are by land-use strata. For single-variable regression a combined estimate for several strata may be used. Estimates are by analysis districts.

### 4.15. Full Scene Classification and Aggregation

Once a statistics file has been made containing categories for all covers of interest, the full Landsat MSS scene is classified. The classification is done on the CRAY. Since the classification procedure is well-suited to the vector architecture of the CRAY, and also to that of the other commercially available supercomputer, the CYBER-205, full-scene classification is quite economical. Also, due to the high capacity of supercomputers, it is possible to process many scenes in a short period of time.

Aggregation is the process of tabulating the categorized file by category and stratum. Aggregation does not lend itself to vectorization. It is, however, very simple and does not use much CPU time. Therefore, it is far more economical to do the aggregation on the CRAY, including transfer to

the CRAY of the relatively small strata mask files and transfer from the CRAY of the relatively small aggregation files, than would be the case if the large full-scene classified file were transferred from the CRAY for aggregation at BBN or elsewhere.

### 4.16. Large-Scale Estimation

Large-scale estimation computes the estimate, variance, and other values by stratum and analysis district. Inputs include the estimator parameter file from sample estimation, the aggregation files for all counties in the analysis district, and the frame unit file. The type of estimation, single variable regression or multi-variable regression, is determined in sample estimation and is stored in the estimator parameter file. The results of estimation are displayed as well as being stored in the estimator results file for use in the accumulate-estimates program.

### 4.17. Accumulate Estimates

The accumulate-estimates program is used to create tables of the various estimates from the estimation parameter file and also to compute totals as requested by the user. For areas not covered by Landsat MSS, direct expansion estimates are computed. Accumulation requires the frame unit file and the estimator results files as inputs. The only outputs are printed listings.

## 5. SUPERCOMPUTER USAGE

Using Landsat MSS data to compute estimates for entire states requires that large amounts of data be processed. Also, since estimates must be delivered late in the year, all the data must be processed in a relatively short time. This has led SRS to take advantage of the fast processing times and generally superior handling of large data sets associated with supercomputers [10]. The term supercomputer will be here taken to mean a high speed vector processing machine with one instruction stream and many data streams, often referred to as SIMD, Single Instruction Multiple Data. A Landsat MSS scene consists of several million data elements, the pixels, all of identical format. Algorithms such as maximum-likelihood classification which process the pixels independently are ideal for such machines, since a number of pixels may be processed simultaneously. Other algorithms, such as CLASSY, in which there are more dependencies between the pixels will not work so well, but even these may have areas which can be optimized for supercomputer architecture.

Besides high speed processing, a supercomputer which is to be used to process full scenes of Landsat MSS data must have superior data handling capabilities. This means large-capacity disks and high speed channels for transfer of files between tape or other archival storage and disk and between disk and main memory. A large main memory is also very desirable since individual pixels are processed quite rapidly making large buffers very useful.

Currently, there are two commercially available supercomputers, the CRAY and the CYBER-205 from Control Data. Both come in varying configuration. Most sites having one of these machines have recognized the

need for handling large amounts of data and have acted accordingly. This, unfortunately, has not been the case with some experimental supercomputers. The configuration of the ILLIAC-IV, an early supercomputer and the first used by SRS, is a case in point. The inadequate data handling facilities and lack of sufficient main memory caused Landsat MSS processing to be much more difficult than it has since become with the CRAY.

Currently, maximum-likelihood classification, aggregation, CLASSY clustering, and block correlation are implemented on the CRAY. Block correlation is an algorithm for creating a multitemporal scene by positioning one scene relative to another covering the same area on the earth's surface. The positioning is done byperforming a correlation of blocks from one scene into those of another at various shifts and choosing the shift with the highest correlation value. Thus, block correlation registers one scene to another and is used to create multitemporal scenes.

## 6. COMPUTER CONFIGURATION FOR EDITOR PROCESSING

The current computer configuration for EDITOR is centered in the DEC-20 at BBN. BBN is connected to the VAX front end of the CRAY at NASA-Ames through ARPANET. This allows for fairly high speed data transmission of large amounts of Landsat MSS data but is not adequate for full scenes which are mailed on tapes. The PDP-11, used for video digitization, and the microcomputers used for manual digitization are connected to BBN by phone lines. This type of connection allows only slow file transfer, but is adequate for the small files involved.

With the introduction of PEDITOR, various EDITOR programs will be moved to other machines. If full-scene classifications are required, one should try to ensure that a reasonable network connection is maintained to some supercomputer or at least large mainframe, unless the total processing is to be a few scenes widely distributed in time. Otherwise, severe delays are likely to result even though the machine is available at night or on weekends to do full-scene processing.

Another consideration to keep in mind is the available file system on the various machines. EDITOR processing requires several files for each segment, namely the ground data, segment network and mask files, and also Landsat MSS data files and other files. The files are given standard names. If certain systems, particularly some microcomputers, have inadequate file systems or perhaps very restrictive file names, it may be necessary to consider using a data base or at least an aggregation of files so that, for example, all segment mask files for a state are placed in one large file and accessed by special subroutines.

## 7. CONCLUSION

EDITOR has provided a useful tool for computing crop acreage estimates using Landsat MSS data. It has envolved over a long period of time and, hopefully, will continue to do so as new algorithms are tested. The technology is changing rapidly. New types of satellite data, as from Thematic Mapper and SPOT, are becoming available. Also, new computers of all sizes are constantly being introduced. To remain a viable system, EDITOR will have to be adapted to these various changes.

The simple, loose structure of individual programs communicating only by reading or writing files has allowed new programs to easily be added. Also, it has allowed users who may not be interested in estimation to use only parts of the system.

Supercomputers have been an important tool in SRS processing of full states. Although "ordinary computers" are becoming more powerful, the computation load is becoming greater due to new higher resolution satellite data, increasing interest in use of multitemporal data, and more complex analysis algorithms.

## 8. ACKNOWLEDGEMENT

The author would like to thank the many people who have written programs for the EDITOR system and who have contributed ideas for new programs and improvements to existing programs.

## REFERENCES


1. Phillips, T.L. (ed.), LARSYS User's Manual, Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana, 1973.

2. Hord, R. Michael, The ILLIAC-IV, The First Supercomputer, Computer Science Press, Rockville, Maryland, 1982.

3. CRAY-1 S Series Reference Manual, Cray Research, Inc., Mendota Heights, Minnesota, 1981.

4. Houseman, Earl, "Area Frame Sampling in Agriculture," Report SRS-20, Statistical Reporting Service, United States Department of Agriculture, Washington, D.C., November, 1975.

5. Sigman, R., G. Hanuschak, M. Craig, P. Cook, M. Cardenas, "The Use of Regression Estimators With Landsat MSS and Probability Ground Sample Data," Survey Section Proceedings, American Statistical Association Meeting, San Diego, California, 1978.

6. Sigman, R., C. Gleason, G. Hanuschak, R. Starbuck, "Stratified Acreage Estimates in the Illinois Crop Acreage Experiment," Symposium Proceedings, Machine Processing of Remotely Sensed Data, Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana, 1977, pp. 80-90.

7. Ozga, M., and R. Sigman, "An Autodigitizing Procedure for Ground Data Labelling of Landsat MSS Pixels," Proceedings of the Fifteenth International Symposium on Remote Sensing of Environment, Ann Arbor, Michigan, 1981, pp. 1265-1273.

8. Ball, G.H., D.J. Hall, "A Clustering Technique for Summarizing Multivariate Data," Behavioral Science, Vol. 12, March, 1967, pp. 153-155.

9. Lennington, R.K., and M.E. Rossback, "CLASSY - An Adaptive Maximum Likelihood Clustering Algorithm," Proceedings of the LACIE Symposium, NASA Manned Spacecraft Center, Houston, Texas, 1982.

10. Ozga, Martin, "Experience With The Use of Supercomputers to Process Landsat MSS Data," Symposium Proceedings, Machine Processing of Remotely Sensed Data, Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana, 1984, pp. 276-280.